New Entries to the SPL EDICS for Audio and Acoustic Signal Processing

Christensen, Mads Græsbøll; Rabenstein, Rudolf

Published in:
I E E E Signal Processing Letters

DOI (link to publication from Publisher):
10.1109/LSP.2012.2224516

Publication date:
2012

Document Version
Early version, also known as pre-print

New Entries to the SPL EDICS for Audio and Acoustic Signal Processing

Mads Græsbøll Christensen, *Senior Member, IEEE,* Rudolf Rabenstein, *Member, IEEE*

***Abstract*—This letter describes some of the new entries to the Signal Processing Letters (SPL) Editors Information Classification Scheme (EDICS) for the topic Audio and Acoustic Signal Processing.**

I. Introduction

The Editors Information Classification Scheme (EDICS) for the IEEE Signal Processing Letters (SPL) in the previous version from the year 2005 [1] contained only one entry for audio: *AEA-AUEA Audio and Electroacoustics*. While this term still breathes the air of vacuum tube amplifiers and speaker cabinets, the recent years have brought new processors with even higher computing power, data storage capacities beyond the terabyte range, higher data rates for internet up- and download, and a variety of different kinds of audio interfaces. All these developements come at ever falling prices and new audio-related products penetrate the market quickly.

In the past, the term *electroacoustics* described a technology for recording studios, stage reproduction and cinemas. But today audio signal processing is found everywhere, in home theatres, home and mobile recording, mobile devices, in passenger cars, and in advanced hearing aids. Traditional audio storage media are being replaced by internet access to cloud storage. The technology for spatial recording, reproduction, and for the analysis of the acoustic environment is advancing. Research at the interface between hearing system and brain provides computational models for human sound perception.

To cope with the emerging diversity of signal processing applications in audio, the recent update of the SPL EDICS from September 2012 [2] includes now a number of new entries. Some of these are described below.

II. EDICS for Audio and Acoustic Signal Processing

The following subsections introduce some of the new subtopics for Audio and Acoustic Signal Processing. Only a rough categorization of the main activities, current and future challenges as well as future possibilities are given. Since no account of past achievements is intended, there are no references to individual contributions.

The naming scheme is reminiscent of the previous designation *Audio and Electroacoustics (AEA)*. The abbreviation AEA has been kept for compatibility; it stands now for *Audio and Acoustic Signal Processing*.

A. AEA-RES: Audio and Speech Signal Restoration

Audio and speech signal restoration deals with restoring, or generally improving, the quality of recorded speech and audio signals. Over time, the quality of signals stored on gramophone records or analog tapes may have degraded in various ways, and imperfections in equipment may have introduced annoying artifacts, like clicks, hiss, crackle or buzz in various steps of the recording process. Moreover, parts of recordings may be missing, due to, for example, the medium being physically broken, and this problem has become no less relevant with the usage of block-based coding techniques. Restoration aims at removing these artifacts without destroying the original contents.

A well-known application of audio restoration is in forensics, a prominent example being the audio recordings of the assassination of President J. F. Kennedy, which some hoped could help shed light on how many shots were fired. Other important applications include the preservation of important historic recordings, and remastering of old music recordings. Many national broadcasting agencies have digitized their entire archives, and so have many libraries, and most music is today distributed in a digital form.

Depending of the exact type of artifact to be removed, various approaches have been pursued. Removal of additive noise of a stochastic nature using linear filtering is a fairly general and well-studied problem whose roots go back as far as the work of Norbert Wiener. Other problems and the approaches used to deal with them are more specific to speech and audio signals and the way they have traditionally been recorded and stored. Clicks are, for example, typically dealt with in a two-step fashion where first the location of the corrupted samples are located, whereafter they are replaced by estimates obtained using a model of the audio signal. Some of the most successful methods have been based on models like auto-regressive (AR), auto-regressive moving average (ARMA), or sinusoidal models. These models have been used to estimate corrupted samples using, for example, interpolation in the parameter domain or using a Bayesian framework.

In the past decade, a number of things have happened that may lead to new advances in speech and audio signal restoration. Firstly, Voice over IP (VoIP) has become prevalent. Some of the problems encountered in transmission of speech over the Internet are similar to those encountered in restoration, namely that parts of the signal may be missing or corrupted. In VoIP, the problem of replacing such parts is referred to as packet-loss concealment, and many new methods have surfaced for dealing with this problem, and it is possible that they can serve as inspiration for new methods for restoration (or vice versa). Secondly, it has been (re-)discovered that signals can be sampled below the Nyquist rate, provided that they are sparse in some domain. This has led to the principle known as compressed sensing. In compressed sensing, sampling is

performed by forming (possibly random) linear combinations of the signal to be sampled. The signal can then be reconstructed (under certain conditions) using sparse approximation methods. This has led to a flurry of activity in finding new methods for sparse approximations. A special case of such a sampling process is one where random samples are lost, and, hence, it is possible, in principle, to apply sparse approximation techniques to restoration problems. Thirdly, quite some progress has been made in estimation theory regarding methods for parameter estimation and spectral estimation for missing data, but many of these have yet to be adapted and applied to speech and audio signals.

B. AEA-AUD: Auditory Models and Anthropomorphic Processing

Auditory models seek to capture the effective processing of the human auditory system, or parts thereof. Such models have many applications including studying the human auditory system, assessment of audio quality and speech intelligibility, or online or offline optimization of audio processing systems.

In the past decade of audio coding research, the auditory models developed have provided simple metrics that reflect the perceptual consequence of audio coding errors. Early measures took only spectral masking phenomena into account while more recent ones also account for, at least some, temporal phenomena. The end result has been that a much higher quality can be obtained at the same bit-rate as compared to when mean square error based measures are used. While such metrics are often only approximate and only account for some phenomena, they allow for flexible schemes that can automatically adapt to changing signals or transmission conditions, and it is hard to argue against the sucess audio coding has seen in the past 15 years.

Auditory models may also serve as inspiration for how to solve audio processing problems. In some cases, the objective of audio processing is exactly to mimic the auditory system, in others the auditory system provides *a* solution and, hence, one possible solution would be one that does the same. Processing that mimics the auditory system or its behavior is often referred to as being anthropomorphic. In many cases, we are still strugling to achieve performances similar to those of the auditory system with our clever algorithms. For example, it has proven illusive to achieve increases in speech intelligibility via single-channel speech enhancement and separation, despite the many advances we have witnessed in these fields within the past few years. Yet, a danger lies in blindly mimicking the human auditory system or adapting models of parts thereof and integrating them into algorithms. The combination of a statistically founded estimator with a perceptually motivated pre-processor, for example, does not necessarily makes sense; the optimal pre-processor is the one that conditions the signal to the underlying assumption of the estimator.

Auditory models have the potential to lead to mathematically tractable definitions of signal processing problems, as has been the case in audio coding. If this can be done for other problems as well, for example for speech enhancement and separation, then it would, arguably, be possible to achieve both increased quality and intelligibility. An obstacle is, of course, that in many applications, unlike in audio coding, no reference signal is available. A promising sign is that quite some progress has been made recently in determining measures of speech intelligibility. However, for these models to be useful in this context, the auditory models must be in the form of metrics to be mathematically tractable, and this is where the main challenge lies. Then, it will be evident whether the optimal solution to the resulting problem turns out to be an anthropomorphic one.

C. AEA-MIR: Content-based Processing and Music Information Retrieval

Content-based processing is processing of or based on the *contents* of signals or representations thereof, here specifically in the context of music. Music information retrieval (MIR) deals with the problem of extracting useful information from music signals. As such, it requires both strong knowledge of signal processing and the nature of music signals (ranging from the physics of musical instruments to cultural aspects) and is, hence, a multi-disciplinary field. The past decade has seen this field rise from relative obscurity to prominence with entire conferences, sessions at conferences and special issues of journals being dedicated to the topic. Recent efforts have focused on problems such as music recommendation, polyphonic transcription, cover song identification, genre classificiation, mood classification, artist or composer identification, key detection, audio to score alignment, melody extraction and many more. Sometimes these problems are not so much of interest in themselves. Rather, they form a basis for testing and comparing algorithms that seek to solve some underlying problem, like measuring music similarity.

Many of the aforementioned problems have formed the basis of tasks (or competitions as some appear to view them) in the Music Information Retrieval Evaluation Exchange (MIREX). In these tasks, data sets are made available for training and algorithms are submitted for evaluation, and results and rankings are then later published by the organizers. This has contributed greatly to the definition of relevant problems, a standardized way of testing algorithms and reporting results. On the flip side, it has also often meant that the most effort has been put into designing complete systems capable of performing well in these tasks rather than analyzing and solving the underlying scientific problems, many of which are still not well understood. In short, a well-functioning system does not always equal a relevant, novel, scientific contribution to the field.

The field is still very much in its infancy, though, when compared to, for example, speech processing, and faces some important challenges in the years to come. One of the challenges of the field is that several of the problems that researchers have tried to solve are fundamentally ill-posed. A well-posed problem (following J. Hadamard's definition) is one for which a solution exsits, this solutions is known to be unique, and it does not change much when there is a small change in initial conditions (i.e., some kind of continuity applies). When the labeling of training data in a classification

problem is not unique (like, e.g., in genre classification), the underlying problem is ill-posed in this sense. Similarly, when the signal in a segment of a song can be described equally well in several ways, the problem of determining the *true* description is ill-posed. A related challenge is that the problems that researchers seek to solve are not always defined in a mathematically rigorous way, the end result being that it is entirely unclear what problem is actually solved and what the properties of the solution are. In other words, we do not know why and when it works and are at a loss when it does not. Hence, a major challenge in the field is to strive towards more well-defined problems and more rigorous approaches to solving these problems.

Finally, a key challenge is to move from tasks defined by experts and researchers to tasks defined by the other people and organizations who we would like to use the algorithms, be it music enthusiasts, librarians, musicians or online music stores. In many cases, it is probably not even possible to define these tasks beforehand, and we must rather rely on user-interaction. This will ultimately determine whether we are successful or not.

D. AEA-AMS: Audio Analysis, Modification, and Synthesis

Audio analysis, modification and synthesis is generally concerned with the analysis (e.g., spectral estimation, parameter estimation, transformation), modification based on such analysis, and synthesis of audio signals.

Analysis aims at transforming an audio signal into a meaningful parametrization or a transform domain, wherein the signal can more easily be modified (for example, time-stretched, pitch-shifted, or otherwise morphed) and then later resynthesized. Over the past decade, significant progress has been made in new and better methods for audio analysis. In particular, advances have been made in parametric models for audio signals (including models that capture a variety of phenomena, like onsets and vibrato) and methods for finding the parameters of these models, including several new methods based on principles such as Bayesian estimation, subspace methods, optimal filtering etc. Another new methodology that has surfaced quite recently is non-negative matrix factorization (NMF), a principle in which a non-negative data matrix is factored into two matrices whose entries are also constrained to being non-negative. NMF was originally proposed for modeling images, but has proven useful for modeling audio spectrograms, and this has led to many new methods for solving classical audio processing problems, like source separation, music transcription, etc. Sparse approximations and compressed sensing are also topics that have led to many new interesting ideas and methods that may find applications in audio analysis, modification and synthesis. In this connection, it is interesting to note that the history of applying the ideas behind sparse approximations to speech and audio signals actually dates quite far back, at least to the 1980s.

Audio resynthesis from a modified parametric model is used to correct the intonation of a musical instrument or to change the pitch or key of a recording. Modifications of the human voice range from removal of coarseness or adding of vibrato to gender change of a singing voice. However, digital music can also be synthesized from other models than those obtained by previous analysis. Advanced synthesis methods rely on abstract instrument models based on either the physical properties of a real or virtual musical instrument or on circuit diagrams of analog synthesizers or vacuum tube amplifiers.

E. AEA-MUL: Multichannel Audio Processing

Multichannel audio processing is an indispensable tool for other audio processing tasks like spatial audio recording and reproduction (Sec. II-F) and the analysis of acoustic environments (Sec. II-G). In the IEEE Signal Processing Society, multichannel processing is generally represented by the Technical Committee on Sensor Array and Multichannel Processing (TC SAM). Nevertheless, an EDICS entry on this topic under AEA Audio and Acoustic Signal Processing is justified, because multichannel processing for audio applications has its own flavor.

One of the most prominent purposes for multichannel processing is beamforming. In audio recording, microphone array beamforming allows for a flexible handling of variable directivities by steering the maximum sensitivity in the direction of a desired source. Conversely, also a null may be positioned at the direction of a noise source. Advanced techniques estimate these directions on the fly from the multichannel microphone data. Beamforming is also applied to loudspeaker arrays. For music reproduction so-called line arrays distribute the sound power more or less evenly to the audience.

Another typical application of multichannel audio processing is the estimation of the direction of arrival (DOA) of the incoming sound waves of a distant source. There are two major approaches, beamforming and time difference of arrival (TDOA) estimation. The beamforming approach constantly scans all possible directions and looks for directions with high incoming signal level. The TDOA approach estimates the time difference between the channels by correlation methods and then converts time differences to angles of arrival. Advanced methods determine also the source distance.

Microphone arrays may be built in different forms. Mobile technology permits to use the microphones of various mobile devices in the same room as a microphone array. Under these circumstances, the array geometry is never exactly known and may even vary during recording. Here, automatic array calibration as another way of multichannel processing steps in.

Multichannel audio processing is also an issue in the spatial audio reproduction techniques described in II-F. In existing systems, the number of independent loudspeakers ranges from some ten to many hundred and the parallel calculation of the loudspeaker signals from an intermediate representation needs to be performed in real-time at the audio sampling rate. An additional restriction is added for the spatial reproduction of live music where the real-time requirements include also low latency in all reproduction channels.

For different applications of multichannel audio processing, general purpose graphical processing units (GPGPUs) are considered as a new hardware platform. Their processing power is abundant, but latency is a bottleneck.

F. AEA-SAR: Spatial Audio Recording and Reproduction

The classical way of audio recording and reproduction adopts a channel based view. The result of a recording session is prepared for storage or transmission as a number of separate audio tracks, one for each loudspeaker channel. Traditionally, the number of these channels is low: two for two-channel stereo, five plus a low frequency extension for the 5.1 format, and a few more for some cinema formats. At the reproduction site, each of these channels feeds one loudspeaker (possibly two- or three-way) which has to be set up in a standardized way. Here, the number and spatial arrangement of the loudspeakers at the reproduction site has to be considered already while recording. Therefore, this principle becomes unwieldy for higher channnel numbers.

New approaches try to develop intermediate data formats for the representation of spatial sound scenes. These formats shall be independent of the microphone arrangement at the recording site and of the loudspeaker arrangement at the reproduction site. Of course, signal processing capabilities are required to turn the original recording into this intermediate format and again to recover the loudspeaker signals from this format.

The nature of these intermediate format differs between the various approaches. One direction is to expand a recorded sound field into the spatial eigenfunctions of the acoustic wave equation, the so-called spherical harmonics. This approach started as an advanced microphone recording technique under the name *Ambisonics*. Today, dedicated microphone arrays are available that require multichannel real-time signal processing for the conversion into the intermediate format and vice-versa for the synthesis of the loudspeaker signals.

Another direction is based on a paradigm already put forth in the MPEG-4 standardization. The shift in point of view is here from channels to sound objects. A sound object may be a single sound source or a group of sources. The idea is to represent each sound object by its recorded audio data plus some parameters like position, orientation, motion trajectories, etc. This mixture of different data constitutes the intermediate format. Again multichannel audio signal processing is required to compute the loudspeaker signals from the data of each single sound object.

Future research aims at the separation of direct sound and reverberant components of a recording, manipulation of these intermediate formats to reproduce virtual sound fields that are different from the recorded ones, combination with binaural techniques, and at the conversion of vintage channel based recordings for multichannel spatial reproduction.

G. AEA-AAE: Analysis of Acoustic Environments

The main purpose of microphone recordings is to capture a certain source signal: a speaker or singer, musical instruments, animal voices, etc. . Often clean recordings are required, i.e. the recorded signal shall contain as little environmental noise or room reverberation as possible. A certain measure of reverberation is desirable only for some musical instruments or for singing voices. In all cases the influence of the environment is considered as a disturbance (noise) or at best as an amendment of the source signal (reverberation).

However, noise and reverberation signals carry information on the acoustic environment from which they emerge. This information may be retrieved when multichannel microphone recordings are available. They provide either geometric information by extracting the direction of arrival of wavefronts or they allow to exploit the redundancy between the microphone channels for e.g. source separation.

The interest in the analysis of acoustic environments is driven by different applications. In some cases, e.g. for spatial reproduction, the location of a sound source or its motion trajectory is recorded as additional data for later reproduction. When recording speech signals, knowing the location of a specific noise source can help to surpress its contribution to the source signal.

Not only information on sound sources within an enclosure but also on the acoustic environment as such is of interest. This information may consist of global parameters like the reverberation time or the propagation speed of sound waves. Also more detailed information can be retrieved like position, orientation, and the properties of acoustic reflectors. Then the first few image sources in a mirror image source model can be identified immediately from sound recordings.

Future directions include the refinement of reflector localization towards a full geometric model of the enclosure. Here the analysis of acoustic environments meets the intention of acoustic imaging methods, e.g. ultrasound imaging. However the challenge is to work with wavelengths in the audio range, with arbitrary excitation signals like speech, and a low number of microphone channels. In the extreme, forensic applications call for the identification of certain types of acoustic environments from a single-channel recording.

III. Conclusion

The EDICS for Audio and Acoustic Signal Processing include also entries for Feedback and Echo Cancellation (AEA-FEC), Source Separation and Enhancement (AEA-SEP), Audio Coding (AEA-COD), and Audio Processing via Sparse Representations (AEA-SPARSE). These have not been described above since they partly overlap with Machine Learning and Statistical Signal Processing (MLSAS), Speech and Language Processing (SPE), and Sensor Array and Multichannel Signal Processing (SAM).

As we have pointed out, there are many challenges ahead of us and many interesting new ideas as well. We hope that the new EDICS system will be a beneficial tool to share these ideas quickly with the signal processing community.

References

[1] "IEEE Signal Processing Letters Edics," *Signal Processing Letters, IEEE*, vol. 12, no. 5, p. 428, may 2005.

[2] IEEE Signal Processing Letters EDICS, updated September 2012. [Online]. Available: http://www.signalprocessingsociety.org/publications/periodicals/letters/letters-edics/